

DYNAMO: An Algorithm for Dynamic Acoustic Modeling

Françoise Beaufays, Mitch Weintraub, Yochai Konig

Speech Technology and Research Laboratory
SRI International
Menlo Park, CA 94025

ABSTRACT

This paper summarizes part of SRI's effort to improve acoustic modeling in the context of the Large Vocabulary Continuous Speech Recognition (LVCSR) project. It concentrates on two problems that are believed to contribute to the large error rates observed with LVCSR databases: (1) the lack of discriminative power of the speech models in the acoustic space, and (2) the discrepancy between the criterion used to train the models (typically frame-level maximum likelihood) and the task expected from the models (word-level recognition).

We address the first issue by searching for features that help in narrowing the model distributions, and by proposing a neural-network-based architecture to combine these features. The neural networks (NNET) are used in association with a set of large Gaussian mixture models (GMM) whose mixture weights are dynamically estimated by the neural networks, for each frame of incoming data. We call the resulting algorithm DYNAMO, for dynamic acoustic modeling. To address the second problem, we propose two discriminative training criteria, both defined at the sentence level. We report preliminary results with the Spanish Callhome database.

1. Introduction

Many factors contribute to the relatively low performance of state-of-the-art speech recognizers operating on spontaneous, telephone, speech. The diversity of speakers and speaking styles, the typically relaxed articulation, the multitude of pronunciation variants, the presence of extraneous noises, the superposition of more than one voice in some segments, and the distortion due to the communication channel are just a few. Whereas some of these factors can be efficiently dealt with by explicit modeling (*e.g.* vocal tract normalization (*e.g.* [AKC94]), pronunciation modeling (*e.g.* [Slo95, FW97])), many others are left upon the acoustic models whose multi-modal distributions must implicitly model all the remaining variations. This however has the well-known drawback of resulting in broad distributions with overlaps that often cause recognition errors.

In this context, identifying features that act as discriminants in the acoustic space would be useful to narrow the acoustic distributions. If such features can be found, the problem becomes how to use them, and how to ensure that sufficient data sharing is allowed for the model parameters to be reliably estimated. These are the main issues that motivated this work.

In the past decade, contextual linguistic features have been widely used in conjunction with decision tree models, and have significantly improved recognition performance (*e.g.* [BdSG⁺91, YOW94]). Decision trees however make data sharing among different states difficult, and are not well suited to the use of features that are continuous

in nature, as opposed to binary. For these reasons, we chose instead to base our models on neural networks.

More recently, Ostendorf *et al.* [OBB⁺97] showed that a combination of acoustic and prosodic features could greatly help identifying speech segments that were erroneously recognized (32% predictability improvement for a 10-hour training subset of Switchboard). Similar results were reported by various researchers working on confidence measures for word recognition (*e.g.* [WBR⁺97]). Presumably, some of these features, which include various measures of speaking rate, SNR, energy, fundamental frequency, stress pattern, and syllable position, could be directly used to disambiguate large acoustic distributions.

In the field of speaker recognition, the use of handset detectors has dramatically decreased recognition error rates by sorting out carbon button from electret handsets [Rey96, HW97]. The handset type could also be used as an input to the acoustic modeling algorithms.

Another important issue in acoustic modeling is how to capture the dynamics of the speech signal. Much research has recently been devoted to relaxing the independence assumption imposed by most hidden Markov modeling approaches (HMM) and to modeling the correlation between successive frames of data, leading to the family of so-called segment models [ODK96]. Without embarking in this level of complexity, and following a feature-based approach, we propose to include in the acoustic models time features similar to the time index proposed in [GN93, DASW94] and [KM94]. These features don't model correlation but they do alleviate the independence assumption.

Our goal here is to explore the usefulness of such knowledge sources as acoustic discriminants, and to propose an efficient and robust architecture to incorporate them in the acoustic models. Clearly, the richness of the acoustic space representation will have a strong influence on how far this approach can be pushed, but the success of the experiments cited above (handset classification, feature-based error prediction, etc.) indicate that the cepstrum-based representation that most systems use offers enough flexibility for the acoustic models to be significantly improved.

As mentioned before, the architecture we propose relies on neural networks. An important issue related to this choice is the selection of a training criterion to optimize the weights of the networks. The desirable properties for this criterion are (1) to be discriminative, (2) to be closely related to the metric used to evaluate the performance of the recognizer (typically the word error rate (WER)), and (3) to be differentiable with respect to the weights of the neural networks.

Not all the above issues will be discussed in the paper since this

System	Eval '95	Eval '96
baseline	71.00	65.22
+ DT	67.77	64.37
+ CI (size: 1/16 DTs)	68.77	65.22
+ CI (size: 1/8 DTs)	68.27	65.22
+ CI (size: 1/4 DTs)	68.34	65.10
+ CI (size: 1/2 DTs)	67.98	64.49
+ CI (size: 1/1 DTs)	67.69	64.31
N-best error rate	52.54	/

Table 1: N-best list rescoring with decision tree models and context-independent phone models of different sizes: WER in %.

work is still in an early stage. Our first goals were to validate the architecture we propose and to investigate different discriminative training criteria. These two points will be addressed. Feature selection, however, will be the object of future work: for our preliminary experiments, we used a set generic knowledge sources including linguistic features and time indices.

2. Baseline System and Databases

The baseline system for this work is a speaker-independent continuous speech recognition system trained with 75 conversations of Callhome Spanish data and 80 conversations from Callfriend Spanish. It is based on continuous-density, genonic HMMs [DMM96], and uses a multipass recognition strategy [MBDW93] with a vocabulary of 8K words, non-cross-word acoustic models, and a bigram language model. N-best lists are generated, and rescored with the original acoustic models, a trigram language model, and additional acoustic models such as decision-tree-based cross-word models (DT) or large context-independent phone GMMs (CI).

3. Recognition with Large Context-Independent Models

We conducted on the Spanish Callhome database a series of N-best list rescoring experiments with decision tree models and with large context-independent GMMs. The numbers of Gaussians in the GMMs were chosen to be fractions of the numbers of Gaussians used in the corresponding decision tree models. The smallest models had 16 times less Gaussians than the decision tree models, and the largest models had exactly the same size. Recognition experiments were performed with two sets of 200 sentences selected at random from the male evaluation test sets of 1995 and 1996. The results, reported in Table 1, show that, for this database, context-independent models perform as well as or slightly better than decision tree models, provided that the numbers of parameters are equal.

4. The DYNAMO Algorithm

The architecture we propose is based on a hybrid system combining feedforward neural networks and context-independent phone models. Each phone is modeled with a large GMM whose mixture weights are dynamically estimated by a neural network (see Fig. 1), hence the name of the algorithm, DYNAMO. The means and variances of the GMMs are held constant. The inputs to the neural network are the knowledge sources discussed in the introduction. For each data frame, the knowledge sources for each phone are evaluated and inputted into the corresponding NNET. Each NNET outputs a

set of mixture weights, and the likelihood of the observed data is computed from the corresponding phone GMM.

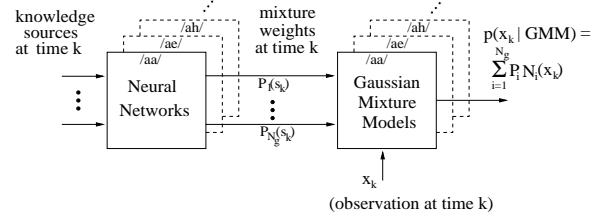


Figure 1: A hybrid NNET-GMM model for dynamic acoustic modeling.

Specifically, the likelihood of an observation, \mathbf{x}_k , with respect to phone φ is given by

$$p(\mathbf{x}_k | \mathcal{N}^\varphi, \mathcal{G}^\varphi) = \sum_{g=1}^{N_g} P_g^\varphi(\mathbf{s}_k^\varphi) N_g^\varphi(\mathbf{x}_k), \quad (1)$$

where \mathcal{N}^φ and \mathcal{G}^φ denote, respectively, the NNET and the GMM associated to phone φ , N_g is the number of Gaussians in \mathcal{G}^φ , $N_g^\varphi(\cdot)$ and $P_g^\varphi(\cdot)$ are, respectively, the g^{th} mixture component and the g^{th} mixture weight in \mathcal{G}^φ , and \mathbf{s}_k^φ represents the vector of knowledge sources for phone φ , at time k .

Because the mixture weights for each phone must sum to one, the training of the neural networks is a constrained optimization problem. To simplify the training procedure, we chose to hard-wire this constraint in the architecture of the neural networks by using a “softmax” output layer [Bri90]:

$$P_g(\mathbf{s}) = \frac{e^{y_g(\mathbf{s})}}{\sum_j e^{y_j(\mathbf{s})}}, \quad (2)$$

where $y_g(\cdot)$ is the g^{th} output of the neural network, before the softmax layer.

The Gaussians in each phone model can be interpreted as a set of basis functions. A multimodal probability density function is then estimated for each observation by taking a linear combination of the basis functions, the weights of which are computed dynamically by the neural network. The discriminative emphasis of certain portions of the acoustic space at each instant has the effect of narrowing the distributions around the acoustic areas where the data is expected to lie.

This architecture thus outputs the likelihoods of the observations. This is in contrast with NNET-HMM hybrids trained for state classification [BM90], where the outputs are state posterior probabilities that need to be converted into likelihoods, and with approaches such as REMAP [BKM95, KBM96] that estimate global posterior probabilities of word sequences.

4.1. Training of the DYNAMO Models

The DYNAMO models are trained in two phases. First, the context-independent phone GMMs are trained with the expectation-maximization (EM) algorithm to maximize the log-likelihood of the training data. The means and variances of these models are retained;

the mixture weights are discarded. Then, the adaptive parameters of the neural networks are trained with the stochastic steepest descent algorithm to optimize some criterion ξ . The neural network weights are thus updated according to

$$\Theta_{n+1}^\varphi = \Theta_n^\varphi + \Delta\Theta_n^\varphi \quad (3)$$

$$\Delta\Theta_n^\varphi = \mu \hat{\nabla}_{\Theta_n^\varphi} \xi_\varphi, \quad (4)$$

where Θ_n^φ denotes the set of neural network weights for phone φ at iteration n , $\hat{\nabla}_{\Theta_n^\varphi} \xi_\varphi$ is the instantaneous gradient of the optimization criterion ξ_φ for phone φ , and μ is a constant that controls the learning rate.

Note that the optimization criterion ξ_φ does not need to be identical to the criterion used to train the GMMs (ML). Indeed, we argue in the next sections that discriminative training is better suited to this task. For now, however, we will assume for simplicity that ξ_φ is the average log-likelihood of the data,

$$\xi_\varphi = \sum_k \log p(\mathbf{x}_k | \mathcal{N}^\varphi, \mathcal{G}^\varphi), \quad (5)$$

where the sum is taken over all the observations \mathbf{x}_k aligned to phone φ .

Applying the chain rule to the derivatives of Eq. 5, and taking Eq. 2 into account, we find

$$\hat{\nabla}_{\Theta_n^\varphi} \xi_\varphi = \sum_j \frac{\partial \xi_\varphi}{\partial y_j} \frac{\partial y_j}{\partial \Theta_n^\varphi}, \quad (6)$$

where

$$\delta_j \triangleq \frac{\partial \xi_\varphi}{\partial y_j} = \frac{P_j^\varphi N_j^\varphi}{\sum_g P_g^\varphi N_g^\varphi} - P_j^\varphi \quad (7)$$

can be backpropagated through the neural network, as in the traditional backpropagation algorithm [RMT86].

Intuitively, the backpropagation term, δ , for Gaussian j is large in absolute value if the posterior probability of the Gaussian is very different from its prior probability P_j , with both probabilities being functions of the knowledge sources for the current data frame.

4.2. Initialization of the neural networks

To fasten the convergence of the neural networks and steer them away from uninteresting local minima, we initially set their weights so that the network outputs are equal to the mixture weights estimated with the EM algorithm.

5. Recognition Experiments with ML-trained Dynamo Models

We performed a set of rescoring experiments with ML-trained DYNAMO models, using linguistic questions and, in some experiments, time features. We chose the linguistic features to be identical to those selected by the decision trees in previous DT-rescoring experiments. The time features for a hypothesized phone aligned to T frames of data were the phone duration, T , and the relative time index t/T , where $t = 0 \dots T - 1$.

Results are given in Table 2, where the baseline obtained by rescoring the N-best lists with the GMMs is given for comparison. These

GMM size	Experiment	WER
$\times 1/16$	no NNETs – baseline	68.77
$\times 1/16$	NNETs w/ ling. feat. & time feat.	69.20
$\times 1/16$	NNETs w/ ling. feat. only	68.92
$\times 1/8$	no NNETs – baseline	68.27
$\times 1/8$	NNETs w/ ling. feat. & time feat.	69.35

Table 2: Rescoring experiments with ML-trained DYNAMO models: WER in %.

numbers show that the introduction of the ML-trained networks increased the overall WER. Further analysis of the results revealed that the likelihood of the test data had increased as a result of training but that the posterior probabilities of the correct models had decreased. This indicated that competing models scored higher than the correct model, which confirmed that discriminative training should be used instead.

6. Discriminative Training Criteria

Discriminative training of speech models was first introduced by Bahl *et al.* under the form of Maximum Mutual Information (MMI) estimation [BBdSM86]. In this framework, the speech models are trained to maximize the mutual information between the observation sequence $\mathcal{X} = [\mathbf{x}_1, \dots, \mathbf{x}_k, \dots, \mathbf{x}_N]$ and the correct word sequence W_c :

$$\Theta^* = \arg \max_{\Theta} I_{\Theta}(W_c, \mathcal{X}), \quad (8)$$

with

$$I_{\Theta}(W_c, \mathcal{X}) = \frac{p(\mathcal{X}, W_c)}{p(\mathcal{X})p(W_c)} = \frac{p(\mathcal{X}|W_c)}{\sum_W p(\mathcal{X}|W)p(W)}, \quad (9)$$

where the sum in the denominator is taken over all possible word sequences, W .

Practical implementations of Eq. 9 for continuous speech recognition include the estimation of the denominator with a phone loop model [Mer88], and its approximation by a sum over the hypotheses in an N-best list [Cho90].

The first optimization criterion we propose is similar to the N-best list implementation of MMI, but differs in that we augment the N-best list with the correct word sequence, W_c . We then maximize the posterior probability of the correct word sequence,

$$P(W_c|\mathcal{X}) = \frac{p(\mathcal{X}|W_c)p(W_c)}{p(\mathcal{X}|W_c)p(W_c) + \sum_{h=1}^{N_h} p(\mathcal{X}|W_h)p(W_h)}, \quad (10)$$

where N_h is the N-best list depth. The inclusion of the joint probability of the observation and the correct word sequence in the denominator makes the criterion depart from the original MMI but has a useful property in terms neural network training as we will show.

Another family of discriminative criteria stems from the motivation of directly optimizing the metric used to evaluate the recognizer, *i.e.* the word error rate. Bahl *et al.* proposed the heuristic “corrective training” procedure in [BBdSM88]. Katagiri *et al.* developed the Generalized Probabilistic Descent method that extends the idea of Bayes optimum classification by introducing smooth classification

error functions, and generalizes this framework to the classification of patterns of variable lengths [KLJ91].

The second criterion we propose consists in minimizing the average number of errors over the N-best list,

$$\text{ANER}(\mathcal{X}) = \frac{1}{N_h} \sum_{h=1}^{N_h} \text{NER}(W_h) P(W_h | \mathcal{X}), \quad (11)$$

where $\text{NER}(W_h)$ denotes the number of errors in the h^{th} hypothesis, and $P(W_h | \mathcal{X})$ is the posterior probability of the h^{th} hypothesis in the (non-augmented) N-best list.

Both criteria are optimized in a stochastic optimization framework, as we will discuss shortly. In both cases, the training procedure requires N-best lists for all the training data. This is typically quite costly but not infeasible, especially if the N-best list depth is limited to a small number of hypotheses (5 or 10).

6.1. Maximizing the posterior probability of the correct sentence

Let $p(i)$ denote the joint probability of a word sequence i (reference or hypothesis) and of the corresponding acoustic sequence,

$$p(i) = p_{LM}(i) p_{AM}(i)^{1/\lambda}, \quad (12)$$

where $p_{LM}(i)$ and $p_{AM}(i)$ are shorthands for the language model and acoustic model probabilities, $p(W_i)$ and $p(\mathcal{X} | W_i)$, respectively, and where λ is the language model weight.

With this notation, we can rewrite the posterior probability of the correct word sequence in Eq. 10 as

$$P(c) = \frac{p(c)}{p(c) + \sum_h p(h)}. \quad (13)$$

Likewise,

$$P(h) = \frac{p(h)}{p(c) + \sum_{h'} p(h')}, \quad (14)$$

denotes the posterior probability of the h^{th} hypothesis in the augmented N-best list. (All posteriors and likelihoods are conditioned upon the set of acoustic models $\{\mathcal{N}_\varphi, \mathcal{G}_\varphi\}$ for $\varphi = 1 \dots N_\varphi$.)

The first training criterion can be expressed as

$$\xi = \frac{1}{N_s} \sum_s \log P_s(c) \quad (15)$$

where N_s is the number of sentences in the training set, and $P_s(c)$ represents the posterior probability of the correct transcription of sentence s .

Adapting the neural network weights according to this criterion amounts to adjusting them after the presentation of each training sentence by an amount proportional to (stochastic gradient update)

$$\nabla \log P_s(c) = \sum_h P_s(h) \left[\nabla \log p_{AM}(c) - \nabla \log p_{AM}(h) \right], \quad (16)$$

where we made use of the property

$$P_s(c) + \sum_{h=1}^{N_h} P_s(h) = 1. \quad (17)$$

Since the acoustic log-likelihoods can be expanded into sums over the observations, \mathbf{x}_k , in the sentence, the above weight update formula modifies the neural network weights only for those frames where the reference and the hypothesis strings do not coincide. In that case, positive training is given to the correct model (c) and negative training is given to the erroneously hypothesized model (h). The log-likelihood gradients $\nabla \log p(\cdot)$ are calculated according to Eqs. 6 and 7. This property results from the fact that the N-best list was augmented with the correct transcription (Eq. 10).

Another desirable feature of this training criterion is that more training is given to hypotheses with high posterior probabilities (the multiplicative term, $P(h)$).

A potential disadvantage is that the correct hypothesis is often not in the N-best list for databases with high error rates. Improving the posterior of the correct sentence may thus result in decreasing the probability of the best (although erroneous) hypothesis in the N-best list.

6.2. Minimizing the average number of errors in the N-best list

The second training criterion we propose is given by

$$\xi = \frac{1}{N_s} \sum_s \text{ANER}_s, \quad (18)$$

where the average number of errors ANER_s in a sentence was defined in Eq. 11.

Note that here the posterior probability of a hypothesis is computed only with respect to the other hypotheses in the N-best list (*i.e.* without taking the reference into account):

$$P_s(h) = \frac{p(h)}{\sum_{h'} p(h')}. \quad (19)$$

Intuitively, minimizing ANER_s “redistributes” the posterior probability mass to favor hypotheses with few errors and penalize hypotheses with more errors.

Again, the weight update formula can be derived by taking the instantaneous gradient of ξ with respect to the weights of the neural networks. The weight update for each sentence is therefore proportional to

$$-\nabla \text{ANER}_s = \sum_h P_s(h) \nabla \log p_{AM}(h) \left[\text{ANER}_s - \text{NER}_s(h) \right]. \quad (20)$$

The characteristics of this weight update formula are quite different from those of the previous criterion. Negative training is given to hypotheses that have a number of errors above average, and positive training is given to hypotheses with a number of errors below

average. Of course, this average, $ANER_s$, evolves with the training of the models. If the learning process progresses correctly, $ANER_s$ decreases with time, thereby progressively decreasing the number of hypotheses that receive positive training. In the limit, all the posteriors $P(h)$ converge to zero except the one that corresponds to the hypothesis with the lowest number of errors, h^* , and $ANER_s$ converges to $NER_s(h^*)$, thereby bringing the training process to an end.

The main disadvantage of this criterion is that positive training is given to all the frames in the best hypothesis, including those associated with incorrectly recognized words. This criterion, however, is closer to the WER metric that we ultimately wish to optimize.

7. Recognition Experiments with Discriminatively Trained Dynamo Models

These experiments were limited to the training of small models (NNETs associated to $GMMs \times 1/16$), with linguistic and time features only.

Fig.2 shows the results of a self-test experiment (*i.e.* the test data is identical to the training data) with the 627 male sentences of the Eval'96 test set of the Spanish Callhome database. The N-best list depth was limited to 10 hypotheses.

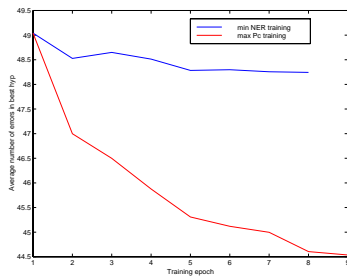


Figure 2: Average number of errors as a function of the training epoch, for both training criteria.

The N-best error rate for this set of sentences was 41.49%. The learning curves show that for the self-test experiment the ANER criterion shows more promise. This however is a cheating experiment, and the generalization properties of the max-posterior criterion may be superior. N-best rescoring of 200 randomly selected male sentences of the Eval'96 test set with the neural networks trained to minimize the ANER gave a significant WER improvement (see Table 3).

models	WER
$GMMs \times 1/16$ — baseline	65.22
min ANER NNETs	63.89

Table 3: N-best rescoring with ANER NNETs, self-test experiment: WER in %.

A fair experiment was conducted with the max-posterior criterion. A set of neural networks was trained from linguistic and time features to output mixture weights for the same small phone models ($GMMs \times 1/16$). The training data consisted of all 15K male sentences in the training set, of which 10 % was held as a cross-validation set. The

models were tested on the same subset of Eval'96 as in the previous experiments. The N-best list depth was limited to 5 hypotheses. The error rate is given in Table 4. The WER improvement is modest but since the phone GMMs in this experiments were small and hence not very detailed, little margin for improvement was left to the NNETs.

models	WER
$GMMs \times 1/16$ — baseline	65.22
max log-post NNETs	64.79

Table 4: N-best rescoring with log-posterior NNETs, fair experiment: WER in %.

8. Conclusions

We described an algorithm to incorporate new knowledge sources in a set of acoustic models, with the objective of dynamically increasing or decreasing the likelihoods of the different modes of the models, thereby narrowing their distributions. The algorithm makes use of feedforward neural networks to dynamically estimate the mixture weights of the speech models, given the knowledge sources for the current data frame.

We argued that the neural networks need to be discriminatively trained, and we proposed two training criteria: maximizing the log-posterior probability of the correct transcription and minimizing the average number of errors in the N-best list.

Preliminary experiments showed a modest but encouraging improvement in WER. We are currently experimenting with larger phone models and increased N-best list depths.

References

- [AKC94] A. Andreou, T. Kamm, and J. Cohen. Experiments in vocal tract normalization. In *Proc. the CAIP Workshop: Frontiers in Speech Recognition II*, 1994.
- [BBdSM86] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1986.
- [BBdSM88] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. A new algorithm for the estimation of hidden markov model parameters. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, New York, NY, April 1988.
- [BdSG⁺91] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny. Context dependent modeling of phones in continuous speech using decision trees. In *DARPA Proc. Speech and Natural Language Workshop*, Pacific Grove, CA, February 1991.
- [BKM95] H. Bourlard, Y. Konig, and N. Morgan. REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities, applications to transition-based connectionist speech recognition. Technical Report TR-94-064, ICSI, Berkeley, CA, March 1995.
- [BM90] H. Bourlard and N. Morgan. A continuous speech recognition system embedding MLP into HMM. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan Kaufmann, 1990.

- [Bri90] J. S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan Kaufmann, 1990.
- [Cho90] Y. L. Chow. Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Albuquerque, NM, 1990.
- [DASW94] L. Deng, M. Aksmanovic, D. Sun, and J. Wu. Speech recognition using hidden markov models with polynomial regression functions as nonstationary states. *IEEE Trans. Speech, Audio Processing*, 2(4), 1994.
- [DMM96] V. V. Digalakis, P. Monaco, and H. Murveit. Genones: Generalized mixture tying in continuous hidden markov model-based speech recognizers. *IEEE Trans. Speech, Audio Processing*, 4(4), July 1996.
- [FW97] M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proc. Eurospeech*, Rhodes, Greece, September 1997.
- [GN93] H. Gish and K. Ng. A segmental speech model with applications to word spotting. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume II, 1993.
- [HW97] L. P. Heck and M. Weintraub. Handset-dependent background models for robust text-independent speaker recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Munich, Germany, April 1997.
- [KBM96] Y. Konig, H. Bourlard, and N. Morgan. REMAP: Experiments with speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Atlanta, GA, May 1996.
- [KLJ91] S. Katagiri, C.-H. Lee, and B.-H. Juang. New discriminative training algorithms based on the generalized probabilistic descent method. In *Proc. Workshop on Neural Networks for Signal Processing*, 1991.
- [KM94] Y. Konig and N. Morgan. Modeling dynamics in connectionist speech recognition - the time index model. In *Proc. Intl. Conf. on Speech and Language Processing*, 1994.
- [MBDW93] H. Murveit, J. Butzberger, V. V. Digalakis, and M. Weintraub. Large-vocabulary dictation using SRI's DECIPHER(TM) speech recognition system: Progressive-search techniques. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages II-319:II-322, April 1993.
- [Mer88] B. Merialdo. Phonetic recognition using hidden markov models and maximum mutual information training. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, New York, NY, April 1988.
- [OBB⁺97] M. Ostendorf, W. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. Technical Report LVCSR Summer Research Workshop, Johns Hopkins U., 1997.
- [ODK96] M. Ostendorf, V. V. Digalakis, and O. A. Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech, Audio Processing*, 4(5), 1996.
- [Rey96] D.A. Reynolds. Mit lincoln laboratory site presentation. In *NIST Speaker Recognition Workshop*, Linthicum Heights, MD, March 1996.
- [RMT86] D.E. Rumelhart, J.L. McClelland, and The PDP Group, editors. *Parallel Distributed Processing*, volume 1. The MIT Press, Cambridge, MA, 1986.
- [Slo95] T. Sloboba. Dictionary learning: Performance through consistency. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1995.
- [WBR⁺97] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke. Neural - network based measures of confidence for word recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Munich, Germany, April 1997.
- [YOW94] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proc. Human Language Technology Workshop*, pages 307-312, Plainsboro, NJ, March 1994.